INDE 597 Team Notepad

Team A

Spring 2024

1 Dynamic Programming

1.1 Exercise 4.1

$q_\pi(11, D)$ is the action value of taking DOWN from state 11, which transitions the agent to the absorbing state. Each transitions earns a reward of -1. Therefore, $q_\pi(11, D) = -1$.

$q_\pi(7, D)$ is the action value of taking DOWN from state 7, which transitions the agent to state 11, earning a reward of -1 in the process. Example 4.1 gives the value of state 11 the equiprobable random policy to be $v_\pi(11) = -14$, so $q_\pi(7, D) = -1 + v_\pi(11) = -14 = -15$.

1.2 Exercise 4.2

If the original transitions remain unchanged–notably, there is no state that transitions into state 15–then the value of state 15 under the equiprobable random policy is given by

$$\begin{aligned}
v_\pi(15) &= -1 + \frac{1}{4}(v_\pi(12) + v_\pi(13) + v_\pi(14) + v_\pi(15)) \\
&= -1 + \frac{1}{4}(-22 - 20 - 14 + v_\pi(15)) \\
&= -1 - 14 + \frac{1}{4}v_\pi(15) \\
\frac{3}{4}v_\pi(15) &= -15 \\
v_\pi(15) &= -20
\end{aligned}$$

The old value of state 13, denoted $v^0_\pi(13)$, is given as $v^0_\pi(13) = -1 + \frac{1}{4}(v^0_\pi(9) + v^0_\pi(12) + v^0_\pi(14) + v^0_\pi(13))$. Observe that the new value of state 13 is given by $v_\pi(13) = -1 + \frac{1}{4}(v_\pi(9) + v_\pi(12) + v_\pi(14) + v_\pi(15)) = v^0_\pi(13) - \frac{1}{4}v^0_\pi(13) + \frac{1}{4}v_\pi(15)$. We know that $v^0_\pi(13) = v_\pi(15) = -20$. Therefore, $v_\pi(13) = v^0_\pi(13)$; the value of state 13 does not change because the DOWN action leads to a state with the same value.

1.3 Exercise 4.3

$$q_\pi(a, s) = \mathbf{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s \ \ \& \ \ \pi_t(s) = a]$$

$$q_\pi(a, s) = \sum_{s', r} [p(s', r | s, a) * (r + \gamma v_\pi(s'))]$$

Being $q_k(s, a)$ the k-th approximation for $q_\pi(s, a)$, we have:

$$\begin{aligned} q_{k+1}(s, a) &= \mathbf{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) | S_t = s \ \ \& \ \ \pi_t(s) = a] \\ &= \sum_{s', r} [p(s', r | s, a) * (r + \gamma v_k(s'))] \end{aligned}$$